111th Academic Year Project Report Competition

Investigation of Intelligent Recognition of Melanoma Based on Systematic Image Feature Extraction

Department and Class: Third Year of the Department of StatisticsName

Student ID：郭依璇 (410978004)(YI-HSUAN KUO)
譚靖蓉(410978011)
謝瑋芸(410978044)
黃名揚(410978056)
李博業(410978058)

Report Date：2023/06/09

Table of Contents

1 Introducion 1

2 Literature Review 2

3 Reserch Method and Steps 3

3.1 Data Source 3

3.2 Data Sampling 3

3.3 Data Preprocessing 4

3.3.1 Hair Removal 4

3.3.2 Removal of Irrelevant Peripheral Areas 5

3.3.3 Lesion Detection 6

3.3.4 Extracting Lesion 7

3.4 Feature Extraction 7

3.4.1 Asymmetry 7

3.4.2 Irregularity 8

3.4.3 Color 8

3.4.4 Texture 8

3.4.5 Proportion of Lesion to Original Image 9

3.5 Feature Normalization 9

3.6 Dimensionality Reduction 9

3.7 Model Development 10

3.7.1 Introduction to Classifier Models 10

3.7.2 Evaluation Metrics 10

3.8 Research Process 11

4. Research Result and Discussion 12

4.1 Impact of Sampling on Results 12

4.2 Training Results of Classifier Models 12

4.2.1 Evaluation Metrics Analysis 12

4.2.2 Feature Importance Analysis 13

4.3 Discussion 15

4.3.1 Results Comparison and Discussion 15

4.3.2 Comparison with Results from Other Studies 17

5. Conclusion 18

6. Appendix 19

A) ABCDE Rule 19

B) hsv Color Three Attributes 20
C) Texture 20
D) Correlation Coefficient Chart of All Variables Before and After Dimension Reduction 23
E) Model Introducion 25

7. References 28

Abstract

The mortality rate of melanoma patients is extremely high, and without early treatment, the effectiveness of radiation and chemotherapy is limited after the spread of melanoma. Early detection and treatment before the spread can significantly increase the chances of recovery. Current identification of this condition primarily utilizes deep convolutional neural networks (CNN), which require numerous images and advanced equipment, and the features are not discernable. This study attempts to directly extract image features, which may aid readers in rapidly locating melanomas in the future.

The research collates publicly available photographs of melanoma and non-melanoma, employs automated cropping, and uses the judgment criteria recommended by Yu Jiarong (2021) [4] to extract features. It employs six classifiers: Support Vector Machine (SVM), Logistic Regression, Random Forest, K-Nearest Neighbors (KNN), Naive Bayes, and Neural Network for melanoma recognition.

The SVM model exhibited the best performance for melanoma recognition and prediction, achieving an accuracy rate of 76.5%. Additionally, among the extracted features, color and texture were identified as the most critical in the recognition process.

1 Introduction

Melanoma is a form of skin cancer that closely resembles moles. According to the American Cancer Society [1], in 2023, there were 97,610 new cases and an expected 7,990 deaths, resulting in a mortality rate of about 0.08%. While this rate may seem low, most skin cancer-related deaths are due to melanoma, highlighting its significant impact on human health. However, its difficult detection often leads to a lack of public awareness.

Current research focuses on using Convolutional Neural Networks (CNN) for melanoma diagnosis. Compared to traditional dermatological assessments based on imagery and experience, CNNs show superior differentiation, significantly outperforming doctors' visual evaluations. Models trained by CNN can assist doctors as a diagnostic aid, but they present some issues, such as high equipment demands, the necessity of large datasets for high accuracy, and the inability of the final CNN model to provide interpretable information to the readers.

Therefore, this study aims to identify accurate statistical and machine models for melanoma recognition under conditions of limited equipment and data. Beyond reducing medical and time costs, it seeks to identify key features for image classification, hoping to eventually support current clinical skin examinations and effectively reduce the clinical assessment burden.

2 Literature Review

Many algorithms for extracting features of melanoma exist in the literature, and the ABCD method is the most frequently mentioned approach. A stands for Asymmetry, B for Border, C for Color, and D for Diameter. Majumder et al. (2019) [3], based on the ABCD rule and segmenting features A, B, and D for processing, utilized an Artificial Neural Network (ANN) model to recognize melanoma with a model accuracy of 98.2%, sensitivity of 98%, and specificity of 98.2%. However, this literature trained the model with only 200 images and tested it on a set of 22 images, without disclosing the source of the download.

Yu Jiarong (2021) [4] proposed a system that, in addition to the ABCD criteria, incorporates E (Enlargement) for the feature of cell expansion. This system can analyze dermatoscopic images of patients' skin from any location, increasing the early detection rate of melanoma, thereby improving the cure rate. The system's feature extraction uses the HSV and RGB color spaces, converts images to grayscale, and applies Gaussian filtering to eliminate noise and hair from the images. The system reportedly achieved an accuracy of 99.20%. Furthermore, the method proposed in this literature can solve problems like noise and hair, greatly aiding in the early detection of melanoma. However, it does not explore how feature values are digitized, lacks a detailed analysis process, and the dataset used contains mole and tumor images that are very clean, clear, high-definition, and do not require any cropping, which does not match real-world conditions.

Gessert and colleagues (2020) [5] utilized data from the ISIC 2019 Skin Lesion Challenge, which included both melanoma and non-melanoma samples, to identify melanomas using a CNN model. However, the best sensitivity achieved on the test set was 74.2%. Despite the use of a large dataset, the final model was uninterpretable. The literature mentions that the ABCDE rule can effectively extract features, but only a few images were used to demonstrate its effectiveness. This study aims to explore whether the ABCDE rule is applicable to the atypical data provided by the ISIC 2019 Skin Lesion Challenge. Given the challenge's data is quite disorganized, this study will first employ methods introduced by Rehman et al. (2022) [6] to address issues with corner frames encountered when capturing dermatoscopic images and how to remove hair from images. The study will first determine the outermost and innermost contours of the corner frames and crop the images accordingly. This allows for the preprocessing of a large number of images from the competition in an automated manner. Then, using more balanced data, seeking additional test sets to avoid experimental errors, and using less perfect images, the same effects can be achieved after preprocessing. The study also practices digitizing feature values and applying them to classifiers to explore identification effects. Ultimately, we hope to achieve the same or even better classification results with less data and standard computer equipment through machine learning and to explain the model results and feature values of melanoma.

3 Research Methods and Steps

3.1 Data Source

This study utilized a variety of datasets for melanoma images: 40 from the PH2 dataset [7], 70 from MED-NODE [8], 162 from the ISIC Challenge [9], 584 from Kaggle [10], and 32,042 non-melanoma images, totaling 856 melanoma images and 32,042 non-melanoma images.

3.2 Data Sampling

A total of 856 melanoma images were collected for this study. The ratio of training to testing datasets was set at 7:3, dividing the melanoma images into 600 for training and 256 for testing. To address data imbalance, over thirty thousand non-melanoma images were similarly split. Random samples of 600 non-melanoma images for training and 256 for testing were drawn. To ensure representativeness, ten sets of 600 non-melanoma images were randomly selected for training, and ten sets of 256 for testing, as shown in Figure 1. Each training set comprised 600 melanoma and 600 non-melanoma images, totaling 1,200 images; each testing set comprised 256 melanoma and 256 non-melanoma images, totaling 512 images. The differences between the ten training and testing sets were compared to demonstrate the representativeness of the non-melanoma image samples.

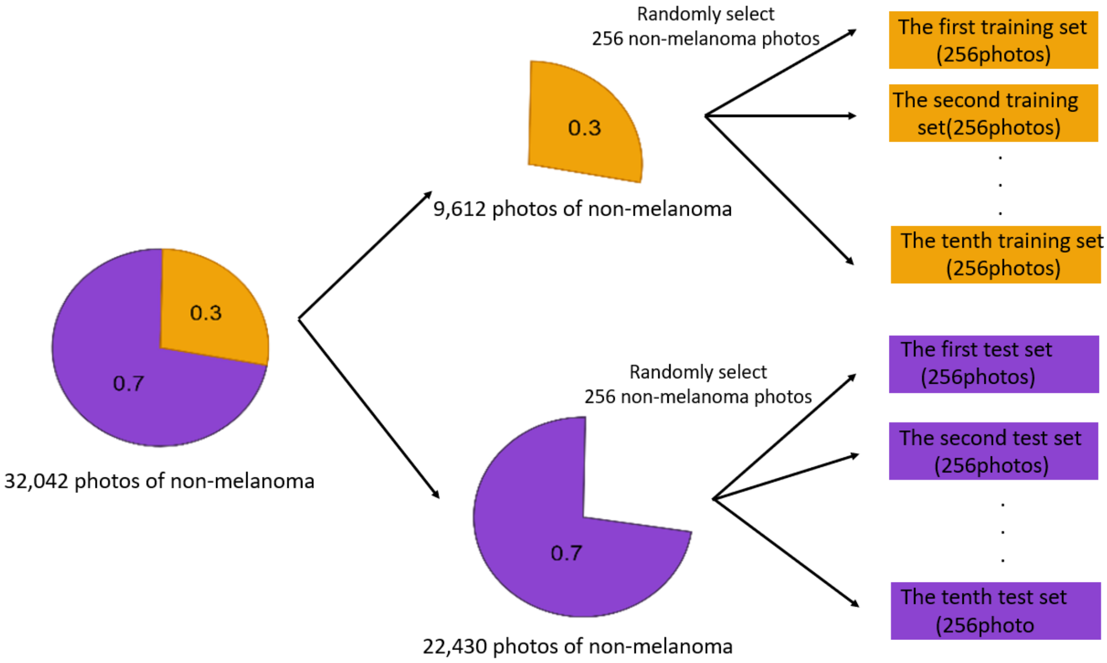


Figure 1: Sampling schematic of the non-melanoma training and test sets

3.3 Data Preprocessing

Due to the issues with the collected original images, such as black and white borders around the images, hair within the images, and other noise, these problems can lead to difficulties in subsequent feature extraction processing, resulting in suboptimal outcomes. Typically, in the field of image processing, opening and closing operations (OpenCV, n.d.) are used to remove hair from images, and images are converted into binary (OpenCV, n.d.) to facilitate the extraction of lesion locations. Therefore, this study processes the images according to the procedure shown in Figure 2, and the following will detail each processing step.

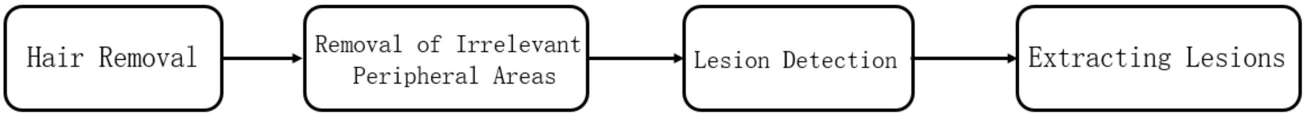


Figure 2: Image preprocessing workflow

3.3.1 Hair Removal

I. Closing Operation

The closing operation involves dilating an image first and then eroding it, which helps fill in small black holes and remove black noise spots. Therefore, we first apply a closing operation to the color image to remove hair, with the results shown in Figure 4, indicating a slight reduction in hair presence.

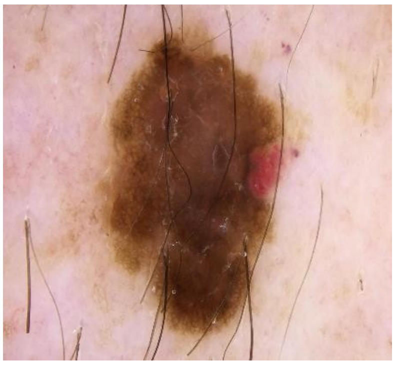


Figure 3: Original Image Before the Closing OperationFigure

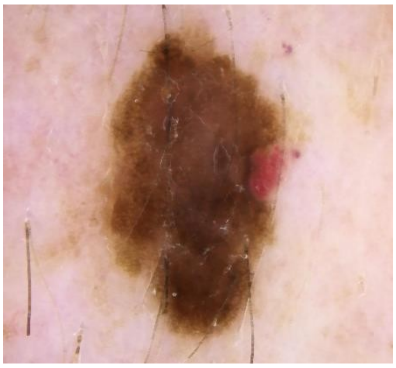


Figure 4: Image After the Closing Operation of Figure 3

II. Opening Operation

The opening operation is applied to images that have already undergone the closing operation, initially eroding and then dilating them. This process removes small white noise and fills in gaps to eliminate tiny bright spots, thereby excluding extraneous noise from outside the lesions. As the opening operation is effective for removing white noise, this study first binarizes the images, turning the residual hair white, and then uses the opening operation to remove it, with the results shown in Figure 6.

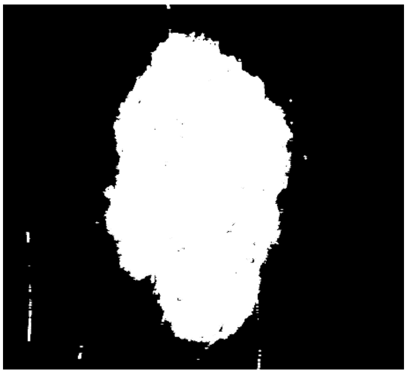


Figure 5: Binarized Image Before the Opening Operation

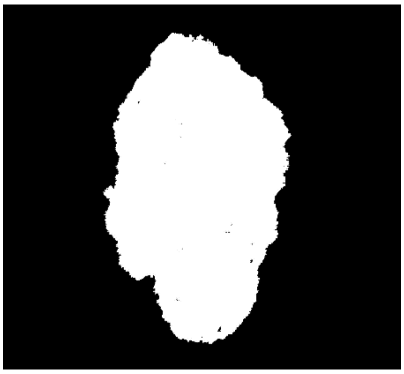


Figure 6: Binarized Image After the Opening Operation

3.3.2 Removal of Irrelevant Peripheral Areas

Referencing Rehman et al. (2022), this approach starts from the center point of the image, using programming to remove peripheral sections irrelevant to this study. For subsequent feature extraction, it is essential to segment the images simply through binarization. First, color images, as shown in Figure 6, are converted to grayscale, and noise is suppressed and smoothed using a Gaussian filter. Then, the Otsu algorithm (Murzova, A. and Seth, S., 2020) is applied to find a threshold: pixels with a gray value above this threshold are turned black, while those below are turned white, with the results displayed in Figure 8.

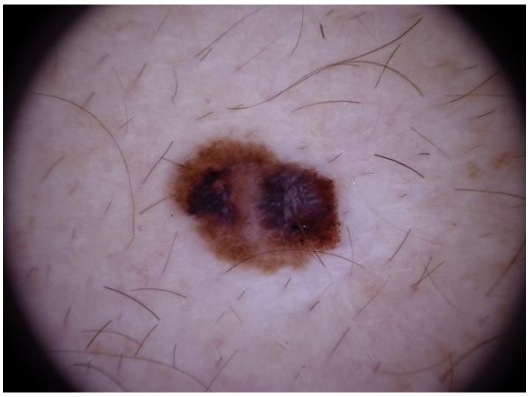


Figure 7: Original Color Image

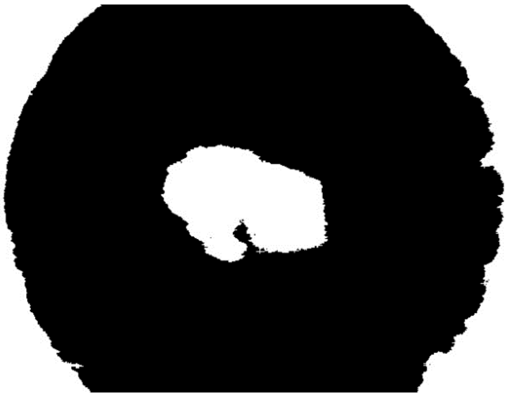


Figure 8: Binarized Image of Figure 7

Due to the presence of black circular frames around some collected images, as shown in Figure 7, which are unnecessary and could affect subsequent feature extraction, this study aims to eliminate such disruptive areas. After binarization, starting from the center point of the original image and moving along the diagonals in all four directions, the study progressively checks pixel values, as illustrated in Figure 9. The intersection points of red and orange dashed lines are identified as the image's border endpoints. Using these points, irrelevant peripheral areas are removed, with results displayed in Figure 10.

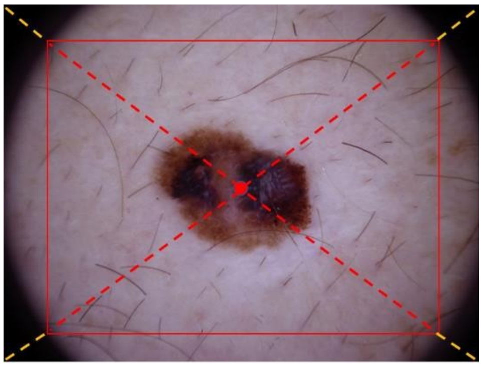


Figure 9: Schematic Diagram of Removing Irrelevant Peripheral Blocks

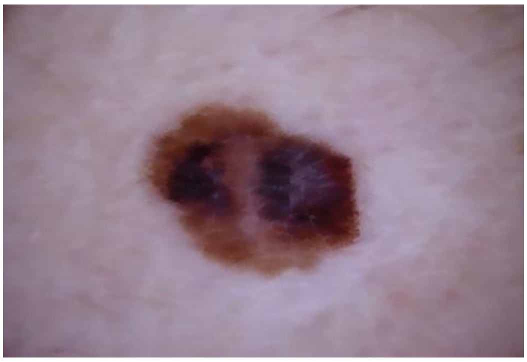


Figure 10: Completed Image of Figure 7 with Irrelevant Peripheral Blocks Removed

3.3.3 Lesion Detection

I. Edge Detection

Even after removing irrelevant peripheral sections, the images might still contain noise that affects feature extraction, as shown in Figure 11. Therefore, edge detection is utilized to outline all contours in the image that could potentially indicate lesions

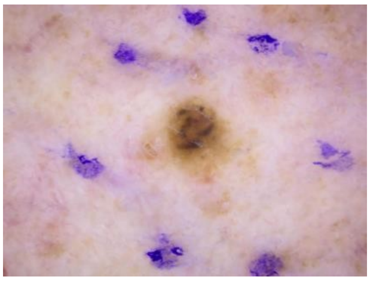


Figure 11: Original Color Image(The lesions are surrounded by noticeable paint.)

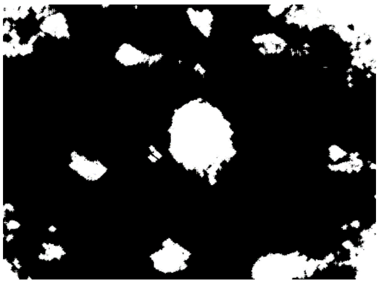


Figure 12: After Binarization of Figure 11

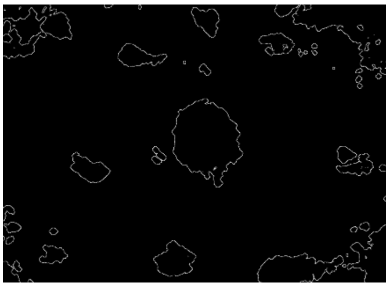


Figure 13: Result of Edge Detection on Figure 12

II. Largest Contour

In the images with removed irrelevant areas, the contours of lesions are typically larger than those of other noise. Thus, this study sorts each contour by length, as shown in Figure 13, and extracts the longest contour, considering it as the lesion contour, as illustrated in Figure 14.

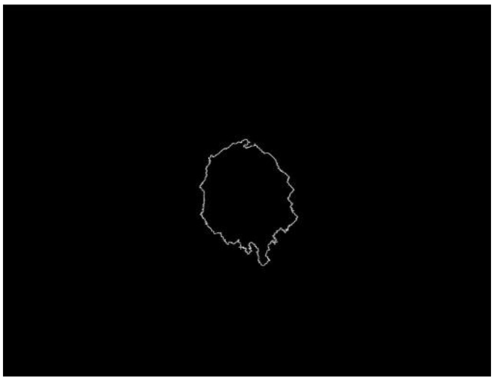


Figure 14: Largest Contour in Figure 13

3.3.2 Extracting Lesions

I. Minimum Bounding Rectangle

Based on the lesion contour, the smallest enclosing rectangle is drawn, as shown in Figure 15. This minimum bounding rectangle is then used as a mask over the original image, as illustrated in Figure 16, and the image is rotated according to the rectangle's angle, as depicted in Figure 17.

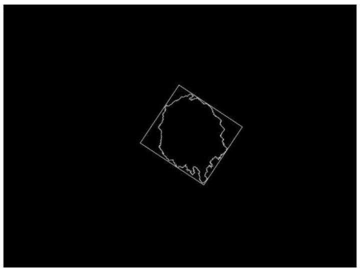


Figure 15: Minimum Enclosing Rectangle around the Lesion Contour

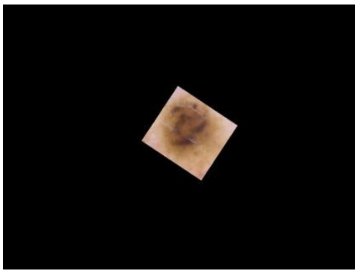


Figure 16: Original Image with the Minimum Rectangle Mask Applied

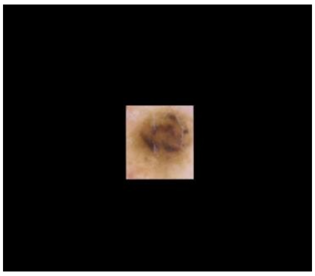


Figure 17: Image 16 Rotated According to the Rectangle's Angle

Ⅱ. Image Cropping

The program is used to crop the rotated rectangular image, extracting the lesion image as the final feature image for subsequent feature extraction, as shown in Figure 18.

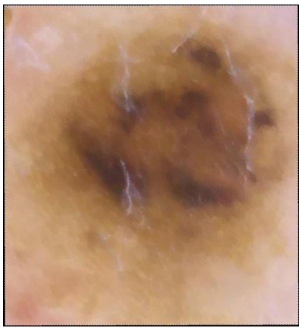


Figure 18: Final Extracted Feature Image

3.4 Feature Extraction

Based on the ABCDE rule and the characteristics of melanoma, this study identifies several directions for feature extraction, including asymmetry, irregularity, color, and texture.

3.4.1 Asymmetry

The image is first binarized and divided into four equal parts, as shown in Figure 19. The area ratio of the lesion in each quarter is calculated, specifically the proportion of the white area in each quarter. Finally, the standard deviation of the four area ratios is calculated and used as the feature value for asymmetry (sym).

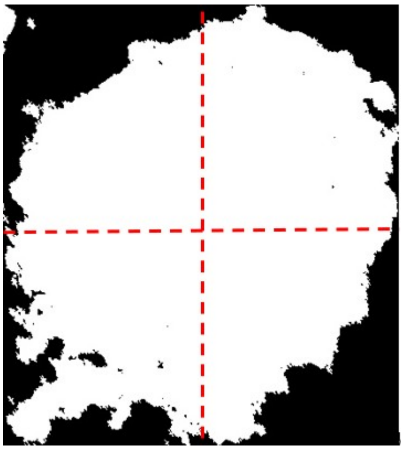


Figure 19: Illustration of Asymmetry Feature Extraction

3.4.2 Irregularity

Based on the irregularity feature, this study outlines a method of drawing a fitted ellipse on the lesion's contour, as shown in Figure 20, and calculates the difference in area between the lesion and the ellipse, divided by the total area of the lesion. The ratio obtained through this method is used as the feature variable for the study's irregularity feature, named the irregularity ratio–ellipse (ae_diff). Another irregularity feature involves drawing the minimum bounding rectangle around the lesion contour, calculating the area difference between the two as a proportion of the lesion area, as shown in Figure 21, named the irregularity ratio–rectangle (ar_diff).

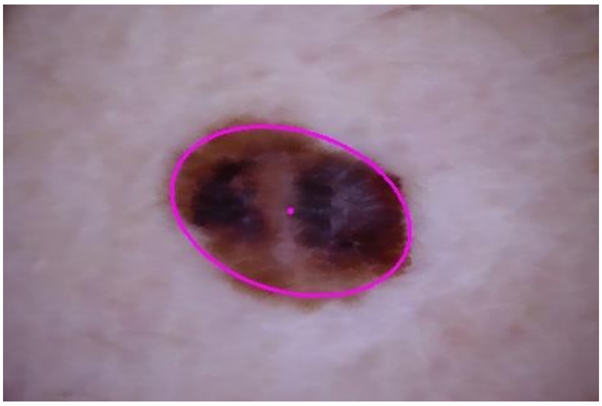


Figure 20: Fitted Ellipse Drawn on the Lesion Contour in Figure 10

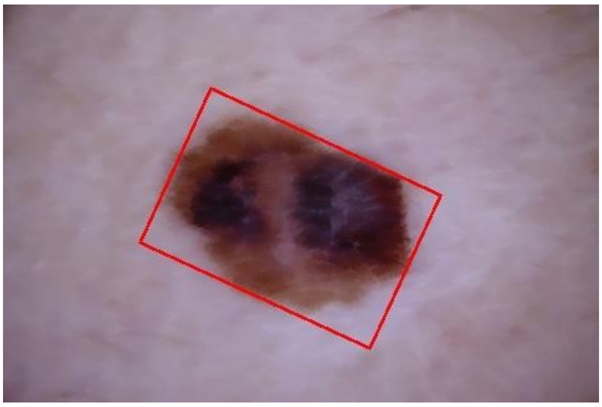


Figure 21: Minimum Bounding Rectangle Drawn on the Lesion Contour in Figure 10

3.4.3 Color

Given the C criterion of the ABCDE rule, melanoma often exhibits uneven color distribution. Accordingly, this study calculates nine features based on the color channels (HSV) of the lesion images, including the mean, standard deviation, and skewness for each channel. These features are the mean hue (h_mean), standard deviation of hue (h_std), skewness of hue (h_skw), mean saturation (s_mean), standard deviation of saturation (s_std), skewness of saturation (s_skw), mean value (v_mean), standard deviation of value (v_std), and skewness of value (v_skw). For details on the three attributes of HSV color, see Appendix B.

3.4.4 Texture

The surface of melanoma is rugged, often presenting a more complex texture than non-melanomas. This study calculates the gray-level co-occurrence matrix (GLCM) of lesion images, yielding five statistical measures at four angles (0 degrees, 45 degrees, 90 degrees, 135 degrees): Contrast, Homogeneity, Dissimilarity, Correlation, and Energy, totaling 20 features. For detailed methods, please see Appendix C.

3.4.5 Proportion of Lesion to Original Image

The feature of proportion (Proportion) takes into account that the area size of the lesion varies in each image. Therefore, this study includes the proportion of the lesion to the total area of the original image as a feature outside of the ABCDE rule.

3.5 Feature Normalization

Figure 22 shows the boxplot of melanoma variables before normalization, revealing significant scale differences between variables, which could impact and introduce errors in subsequent analyses if not normalized. Figure 23 presents the boxplot of melanoma variables after normalization, where the distribution of data shows no significant differences between variables, indicating the suitability of further statistical analysis using different classifiers on the features. ◦

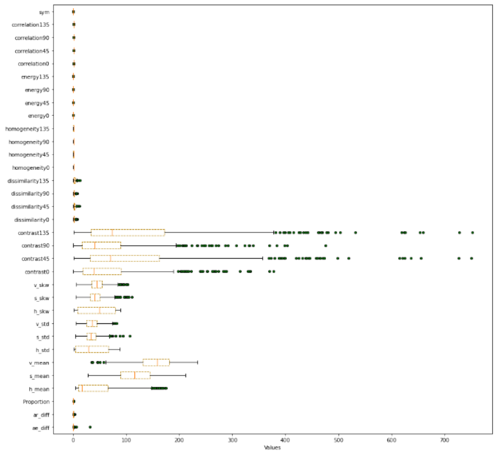


Figure 22: Boxplot of Melanoma Variables Before Normalization

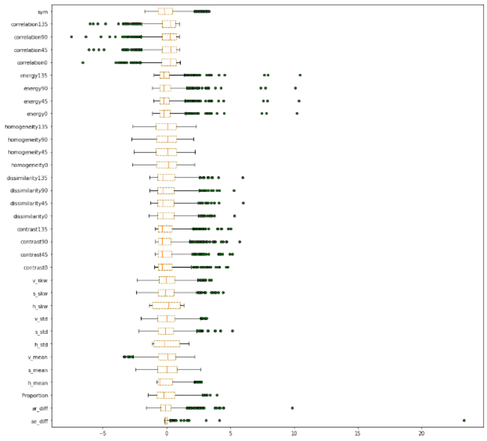


Figure 23: Boxplot of Melanoma Variables After Normalization

3.6 Dimensionality Reduction

Some extracted feature values exhibit high correlation, necessitating dimensionality reduction.

Figure 24 shows that the five variables generated from the gray-level co-occurrence matrix—contrast, dissimilarity, homogeneity, energy, and correlation—exhibit high correlation across the four angles. Therefore, PCA is applied to each to reduce dimensions. After dimensionality reduction, five new variables are generated: contrast (conf1), homogeneity (homof1), dissimilarity (disf1), correlation (corrf1), and energy (energyf1).

Figure 25 demonstrates that the standard deviations and skewness of the three HSV channels are highly correlated, and PCA dimensionality reduction is also applied to them. After dimensionality reduction, three new variables are generated: hue standard deviation skewness (h_ss), saturation standard deviation skewness (s_ss), and value standard deviation skewness (v_ss). For the correlation coefficient diagrams of all variables before and after dimensionality reduction, refer to Appendix D.

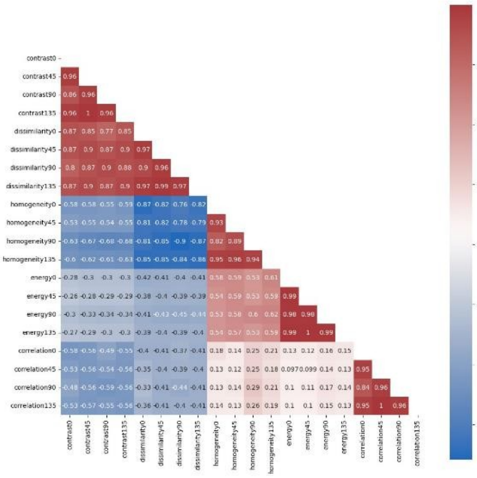


Figure 24: Correlation Coefficient Diagram for Contrast,Dissimilarity, Homogeneity, Energy, and Correlation

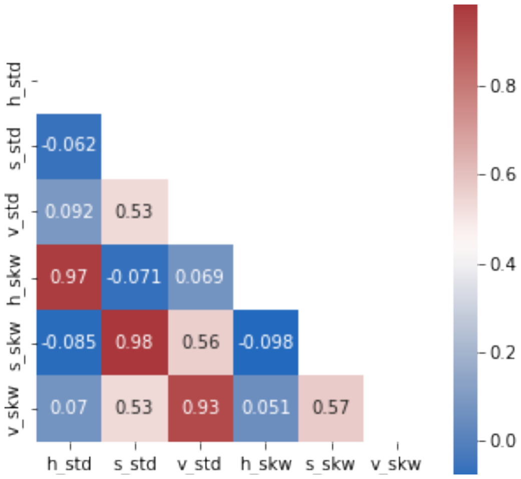


Figure 25: Correlation Coefficient Diagram for Saturation (s) and Value (v) Mean, Standard Deviation, and Skewness

3.7 Model Development

3.7.1 Introduction to Classifier Models

This study employs the following seven classifiers:

1. Support Vector Machine (SVM)
2. Random Forest (RF)
3. Naive Bayes classifier (NB)
4. K-Nearest Neighbor (KNN)
5. Logistic Regression (LR)
6. Neural Network (NN)
7. Convolutional Neural Network (CNN)

For detailed method descriptions, please see Appendix E.

3.7.2 Evaluation Metrics

The evaluation metrics for this study are divided into two aspects : statistical quantitative indicators and program execution time.

I. Statistical Quantitative Indicators

The analysis results of each model will be presented in the form of Table 1, divided into four parts:

(1) True Positive (TP): The case where the lesion is actually melanoma and is predicted to be melanoma.

(2) True Negative (TN): The case where the lesion is actually not melanoma and is predicted to be not melanoma.

(3) False Positive (FP): The case where the lesion is actually not melanoma but is predicted to be melanoma.

(4) False Negative (FN): The case where the lesion is actually melanoma but is predicted to be not melanoma.

Based on these, various indicators to evaluate the statistical method's performance in classifying are constructed. The statistical evaluation indicators are listed in Table 1

Table 1: Analysis Results of Models

		Actual situation	
		+ (Melanoma)	- (Non-melanoma)
Table of Predicted Results	+ (Melanoma)	TP	FP
	- (Non-melanoma)	FN	TN

Table 2: Statistical Evaluation Indicators

Evaluation Metrics	Definition	formula	Indicator significance
Sensitivity	The ratio of cases where the true condition is positive and the prediction result is also positive.	$\frac{TP}{FN + TP}$	Testing the Ability to Correctly Identify Patients with the Disease
Specificity	The ratio of correct predictions to the overall total.	$\frac{TN}{FP + TN}$	Testing the ability to correctly identify individuals without the disease
Accuracy	The ratio of correct predictions to the overall total	$\frac{TP + TN}{TP + FP + TN + FN}$	Testing the ability to correctly identify all patients

II. Program Execution Time

Due to the distinct internal mechanisms of each classifier model, their execution times vary. Therefore, this study also includes time cost as one of the evaluation criteria.

3.8 Research Process

For a complete overview of the research process, please refer to Figure 26.

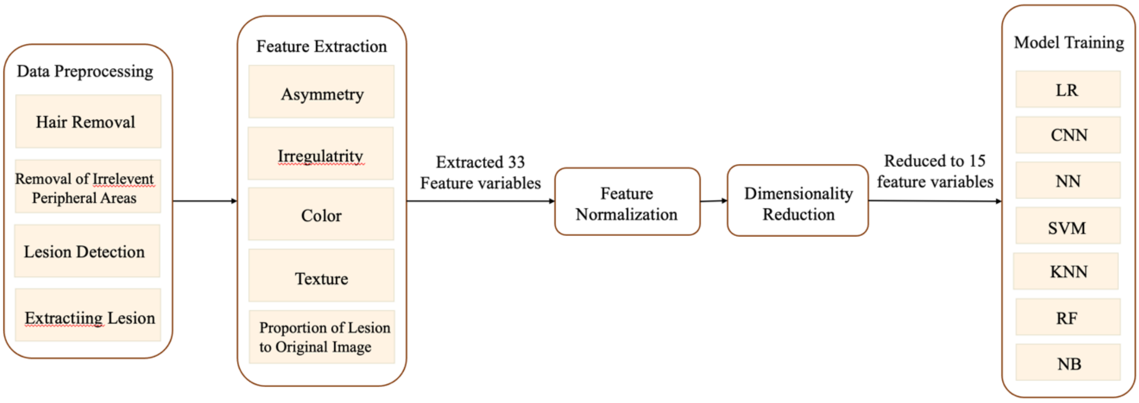


Figure 26: Research Process Diagram

4. Research Results and Discussion

4.1 Impact of Sample Sampling on Results

During sample selection, discrepancies between the sample and the population can arise, impacting the accuracy and representativeness of the research results. This study compares the differences across ten training sets to demonstrate the representativeness of non-melanoma image samples.

By training seven types of classifiers with these ten sets and comparing the outcomes, it was observed from Figure 27 that the accuracy interquartile ranges across the classifiers for the ten training sets were not significantly different. This indicates that the sampling of non-melanoma samples is representative.

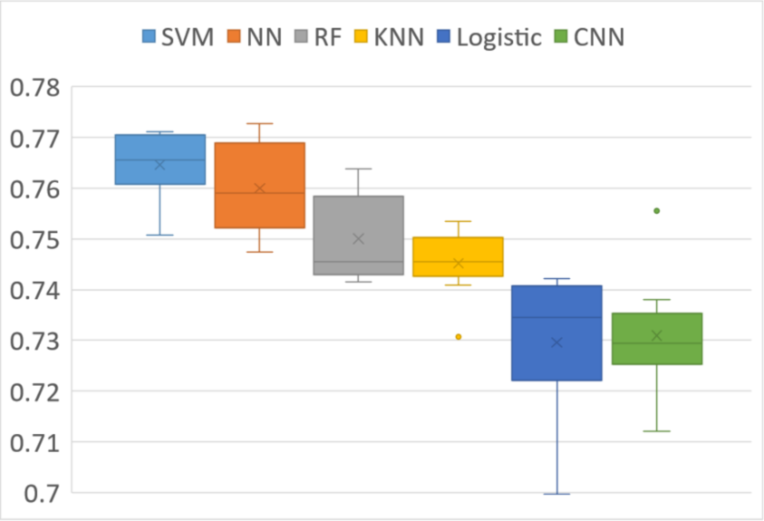


Figure 27: Boxplot of Accuracy for Ten Training Sets After Training with Various Classifiers

4.2 Training Results of Classifier Models

4.1.1 Evaluation Metrics Analysis

This study compares seven types of classifiers. Table 3 presents the average results after training with ten different training sets across various classifiers, and Table 4 shows the average outcomes after each of the ten training sets was tested against ten different test sets. The comparison allows for observations on:

I. Accuracy

The accuracy performance of all seven classifiers was quite good, with Support Vector Machine (SVM), Random Forest, and Neural Network showing particularly excellent performance.

II. Execution time of a program

Apart from NN (Neural Network) and CNN (Convolutional Neural Network), the program execution times of the other five classifiers were all within 0.5 seconds, with the Naive Bayes Classifier having the shortest execution time.

Table 3: Training Results of Classifier Models (Average of 10 Training Sets)

Types of Classifiers	Specificity(%)	Sensitivity (%)	Accuracy (%)	Execution Time (Seconds)
SVM	84.1	84.4	84.2	0.21904
Random Forest	100	100	100	0.60724
Naïve Bayes	45.7	86.5	66.1	0.00090
KNN	100	100	100	0.02534
Logistic Regression	76.2	73.7	75.0	0.00788
NN	100	100	100	12.40170
CNN	100	100	100	32.35060

Table 4: Test Results of Classifier Models (Average of Each of 10 Training Sets Across 10 Test Sets)

Types of Classifiers	Specificity(%)	Sensitivity (%)	Accuracy (%)	Execution Time (Seconds)
SVM	75.8	77.	76.5	0.07623
Random Forest	73.2	76.8	75.0	0.20994
Naïve Bayes	57.5	70.3	63.9	0.00054

KNN	74.7	74.4	74.5	0.00684
Logistic Regression	71.4	74.5	73.0	0.00573
NN	74.6	77.3	76.0	9.14600
CNN	76.6	68.4	72.6	32.35060

4.1.2 Feature Importance Analysis

I. SVM

Using a linear kernel function, SVM can calculate the feature weights. Figure 28 shows the results after taking the absolute values of the feature weights and then sorting them, indicating the importance of each feature in distinguishing melanoma. Red indicates a positive influence, meaning the larger the variable, the higher the probability the image is recognized as melanoma; blue indicates a negative influence, meaning the larger the variable, the lower the probability the image is recognized as melanoma.

From Figure 28, it is determined that the mean value of brightness (v_mean), mean saturation (s_mean), dissimilarity (disf1), and the proportion of the image (Proportion) are important features when using SVM to diagnose melanoma.

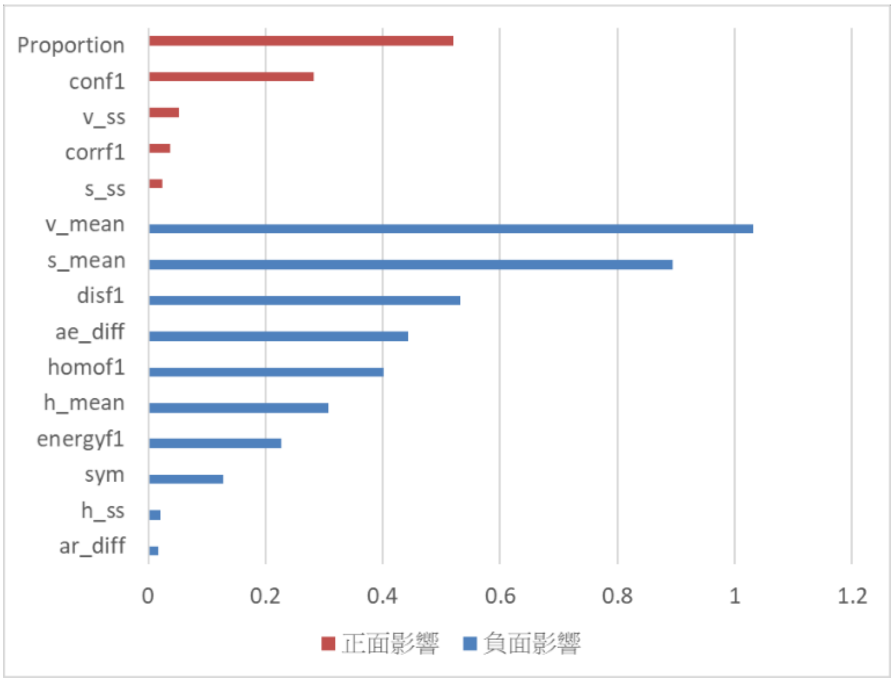


Figure 28: SVM Feature Importance Bar Chart

(The length of the bars in the bar chart represents the average SVM feature coefficients across ten training sets)

II.Random Forest

In Random Forest, feature importance scores are relative, measuring each feature's contribution to the model's predictive capability. Higher Gini importance values indicate a greater impact of the feature on model predictions. Figure 29 shows the average Gini importance values for the ten training sets within Random Forest.

From Figure 29, it is understood that in the Random Forest model, dissimilarity (disf1) and homogeneity (homo1) are important features for diagnosing melanoma using Random Forest.

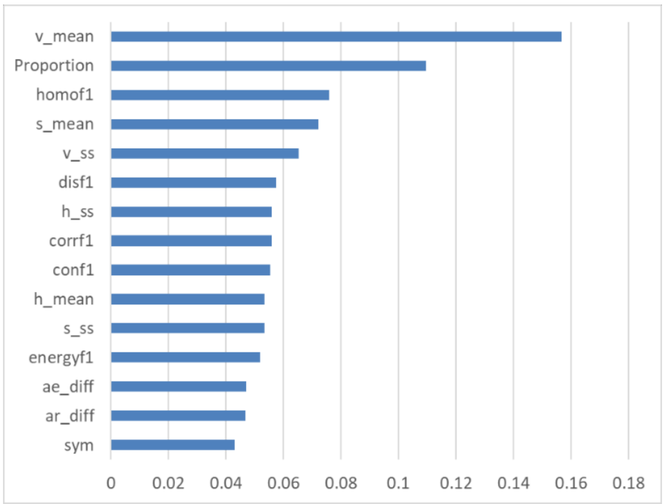


Figure 29: Random Forest Feature Importance Bar Chart(The length of the bars in the bar chart represents the average feature weight values of Random Forest across ten training sets

III. Logistic Regression

Figure 30 shows the standardized average estimates of regression coefficients for each variable across ten training sets, sorted by magnitude to determine their importance in the logistic model, as depicted in Figure 30. Red indicates a positive impact, meaning the larger the variable, the higher the probability the image is identified as melanoma; blue indicates a negative impact, meaning the larger the variable, the lower the probability the image is recognized as melanoma. According to Figure 30, irregularity ratio-ellipse (ae_diff), mean brightness (v_mean), dissimilarity (disf1), and homogeneity (homof1) are significant features when diagnosing melanoma with the logistic model.

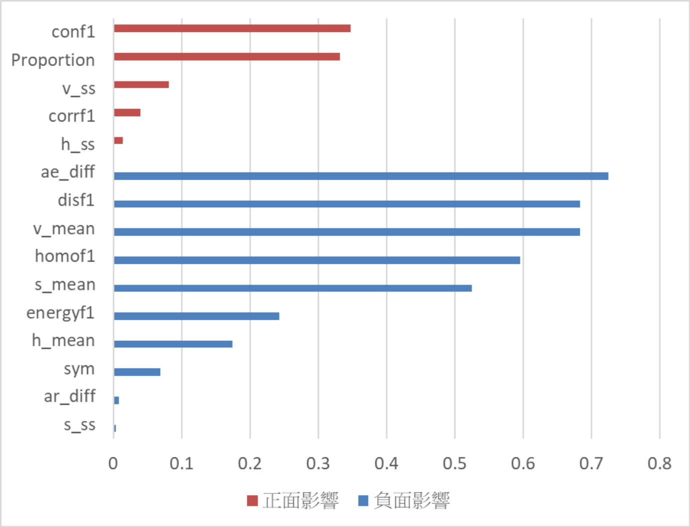


Figure 30: Bar Chart of Average Estimates of Standardized Regression Coefficients for Logistic Regression(The length of the bars in the bar chart represents the average values of the standardized coefficients across ten training sets.)

4.2 Discussion

4.2.1 Results Comparison and Discussion

Regarding classifiers, synthesizing all results, SVM excelled in sensitivity, specificity, accuracy, and program execution time. Although the Naive Bayes classifier had the shortest execution time, its accuracy was the lowest among all classifiers, which might be due to its assumptions not fitting well with the data used in this study. In the classifiers of this study, the importance of various feature variables was evident in SVM, Random Forest, and Logistic Regression models. The key feature variables for each model are summarized in Table 5.

Table 5: Key Feature Variables for Melanoma Identification by Each Classifier

Classifier Types	Key Feature Variables					
	Mean Brightness Value	Mean Saturation Value	Irregularity Ratio - Ellipse	Homogeneity	Proportion of Lesion to Original Image	Irregularity Ratio - Ellipse
SVM	✓	✓	✓	✓	✓	
Random Forest	✓	✓		✓	✓	
Logistic Regression	✓	✓		✓		✓

The discussion will focus on the feature extraction directions mentioned earlier

I.Asymmetry

In model development, asymmetry consistently emerged as a variable of low importance. This might be attributed to the potentially imprecise method of extracting asymmetry features or the complexity and diversity of non-melanoma types, many of which exhibit asymmetrical shapes. Consequently, asymmetry may not be a suitable feature variable for distinguishing melanoma, possibly due to the overlap in asymmetrical shapes between melanomas and various non-melanoma lesions.

II.Irregularity Degree

Irregularity ratio - ellipticity negatively impacts, in model development, it is a variable of high importance, indicating that the more regular the shape, the higher the probability of being identified as melanoma. This is contrary to the expected results of this study. The speculated reason, as mentioned above, is due to the complexity of non-melanoma types, thus making the degree of irregularity potentially not a suitable feature variable for distinguishing melanoma.

III. Color

The average brightness negatively impacts, being a variable of high importance in model development. This indicates that the lower the average brightness, or the closer to black, the higher the probability of being identified as melanoma. Similarly, the average saturation also negatively impacts, being a variable of high importance in model development. This indicates that the lower the average saturation, or the less vivid the color, the higher the probability of being identified as melanoma.

IV. Texture

Homogeneity negatively impacts, being a variable of high importance in model development. This indicates that the lower the homogeneity, or the greater the texture variation and the lower the smoothnes, the higher the probability of being identified as melanoma.。

V. Proportion of Lesion to Original Image

The proportion of lesion to original image positively impacts and is considered an important variable in model development. This means that the higher the lesion proportion, the higher the probability of identifying melanoma, indicating that the proportion of the lesion in the image affects the identification of melanoma. Subsequent research should consider this factor as a control variable.

In summary, the average brightness, average saturation, and homogeneity are feature variables that align with the expected results. Therefore, this study considers color and texture to be key feature variables for identifying melanoma. Asymmetry has low importance across all models, and the degree of irregularity is a feature variable that does not match the expected results. Thus, this study believes asymmetry and the degree of irregularity are less suitable as feature variables for distinguishing melanoma.

4.3.2 Comparison with Results from Other Studies

Most studies on melanoma classification directly apply a large volume of original images into deep learning models for classification, as seen in (Edubirdie, 2022) [14]. However, this approach not only incurs significant medical costs to acquire the images but also requires high-specification computer equipment for model training, otherwise leading to extended execution times. Furthermore, due to the difficulty in interpreting features with deep learning, even if the deep learning model yields good results, it is still challenging to identify clear features for diagnosing melanoma.

Some research uses classifiers other than deep learning models for classification, employing the MED-NODE dataset (Sultana,), and achieves a highest classification accuracy of 77.1% with SVM. However, the images used for training the model in this study are all clear pictures captured by professional equipment, as illustrated in Figure 31. If lower quality images are used for differentiation, as in Figures 3 and 7, the effectiveness of classification is limited.

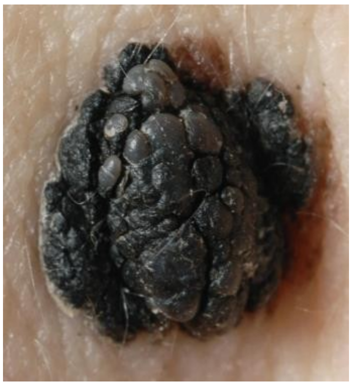


Figure 31: Melanoma Captured by Professional Equipment

In this study, noisy original images were preprocessed to extract lesions, and features were derived based on the ABCDE rule, with classification also performed using SVM, resulting in an accuracy of 76.5%. This approach not only allows for the analysis of images with minor, less distinct lesions and more noise but also reduces medical costs and execution time, and facilitates the interpretation of features to identify key characteristics for distinguishing melanoma. Furthermore, this study addressed data balance and sampling stability, making the results more valuable for reference.

5. Conclusion

After various analyses and attempts, this study incorporated an automated image cropping method to accurately capture the lesion locations in different images and extracted features based on the ABCDE rule, with an interpretation of these features. The highest accuracy achieved using the test set with the SVM model was 76.5%. Although this accuracy does not surpass that of other studies utilizing CNN for classification, this research identified key features for distinguishing melanoma, namely color and texture. This not only narrows down the scope for subsequent melanoma research but also provides the public with clearer criteria for melanoma identification, facilitating early treatment and improving survival rates. Future research could delve further into color and texture to find more related features, increasing the chances of successfully distinguishing melanoma. It is hoped that in the future, automated diagnosis can completely replace current clinical skin examinations, becoming a better basis for detection.

6. Appendix

A) ABCDE Rule

I.Asymmetry

Examine the symmetry of both sides of the feature center. Benign pigmented lesions are often symmetrical, meaning if a line is drawn down the center of the lesion, the left and right halves should be symmetrical.

II.Border

Borders pertain to the feature's edges, observing for any irregular protrusions or indentations. The borders of melanomas mostly exhibit an irregular appearance, as illustrated in Figure 32, whereas the edges of common moles on the skin tend to be smooth, forming circular or elliptical curves, as depicted in Figure 33.

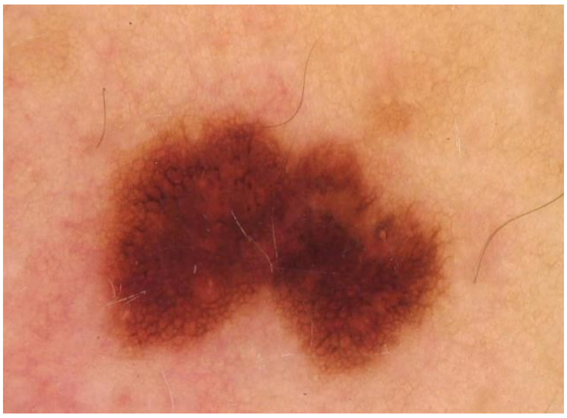


Figure 32: Melanoma with Irregular Borders

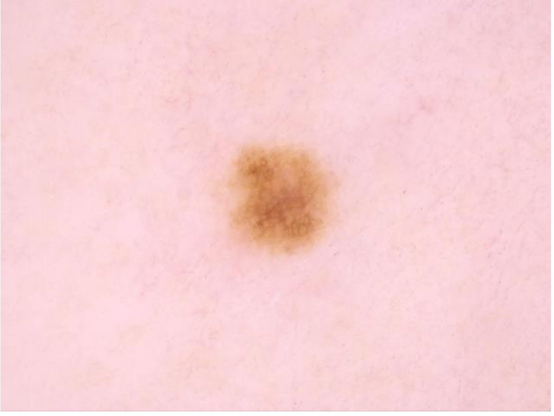


Figure 33: Common Moles on the Skin

III. Color

Color, as a feature, is judged based on whether the color is evenly distributed throughout the feature. Melanomas tend to have a more uneven distribution of color, with blue or red spots easily appearing on a dark brown base. These spots can vary in size and do not have a specific distribution area, as shown in Figure 34. In contrast, the color of moles is mostly evenly distributed, presenting as a solid block of dark brown, as shown in Figure 35.

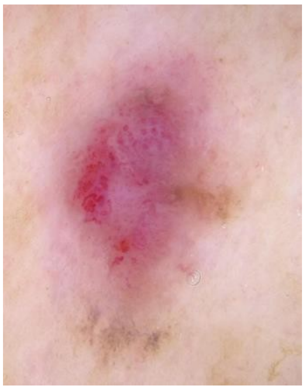


Figure 34: Color Distribution of Melanoma

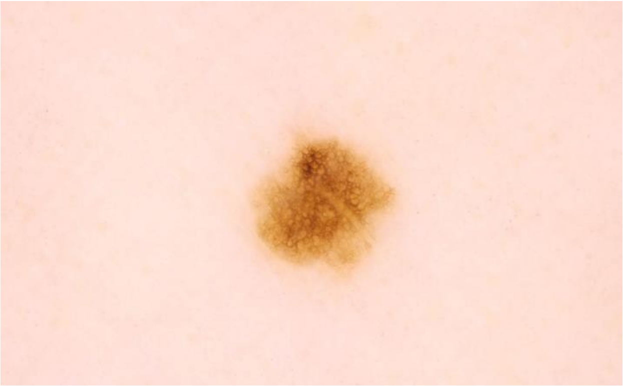


Figure 35: Color Distribution of Moles

IV.Diameter

Refers to the size of the feature's diameter; if greater than 6 mm, there is a high probability of melanoma.

.

V.Enlargement

Changes in the feature, including increase in size, volume expansion, color change, etc.

B) HSV Color Three Attributes

HSV Color Three Attributes Chart, refer to Figure 36.

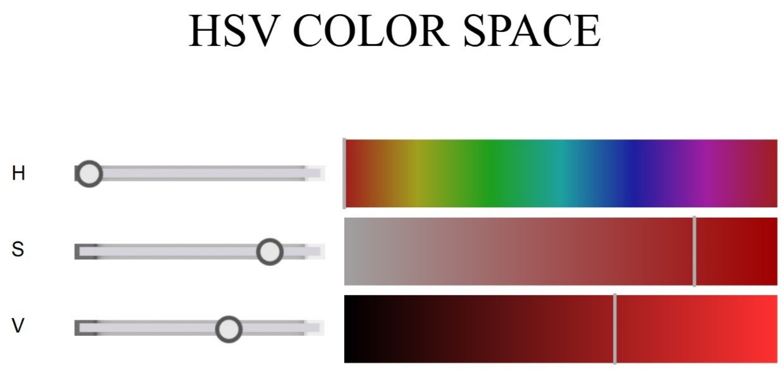


Figure 36: HSV Color Three Attributes Chart

C) Texture

In addition to the uneven color distribution mentioned above, the surface of melanoma is more likely to exhibit irregular elevations and depressions due to pathological changes. As shown in Figure 37, the surface of a melanoma typically has irregular protrusions; conversely, the surface of a common mole is smoother, as illustrated in Figure 38. Both the uneven color and the uneven surface can lead to a more abundant and complex texture on the surface of the lesion. Compared to moles with uniform color and smooth surfaces, the texture of melanomas is generally more pronounced and complex

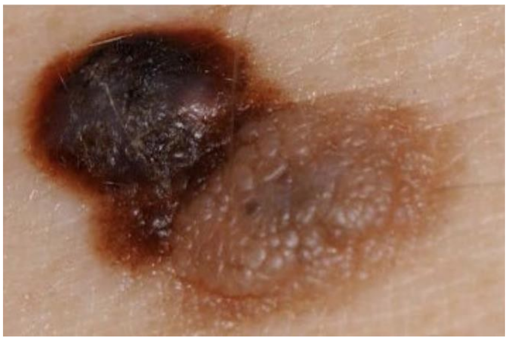


Figure 37: Surface of Melanoma

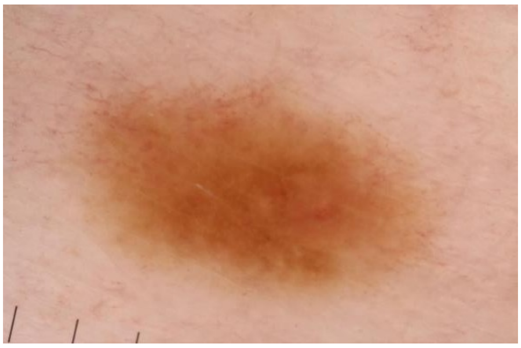


Figure 38: Surface of a Mole

To analyze the distribution of textures, we calculate the Gray Level Co-occurrence Matrix (GLCM) of the image for subsequent analysis. A brief introduction to the GLCM is as follows:

The units on both axes of the Gray Level Co-occurrence Matrix are pixel values, ranging from 0-255 (where 0 represents black, and 255 represents white). The value of GLCM(i, j) represents how many pairs of adjacent pixels in the original color image have the combination of (i, j). Adjacent pairs are calculated in four directions: 0 degrees, 45 degrees, 90 degrees, and 135 degrees, resulting in four GLCMs for each image at different angles.

Moreover, if the original image is converted to a grayscale image and there are many adjacent pixel pairs with the same combination, i.e., (i, i) combinations, the diagonal elements of the GLCM will have larger values. If the grayscale image shows more local variation, the diagonal elements of the GLCM will have smaller values.

Observing Figures 39 and 40, where Figure 40 is a blurred version of the same photo as Figure 39, the blurring causes the texture in Figure 40 to be less clear than in Figure 39. Therefore, the diagonal in Figure 42 is more pronounced than in Figure 41.

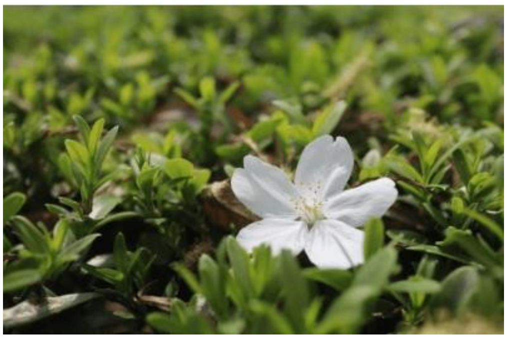


Figure 39: Example of an Original Image

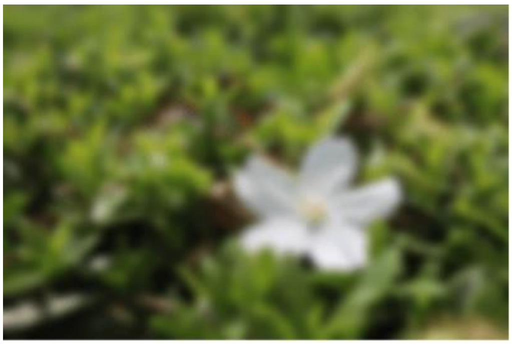
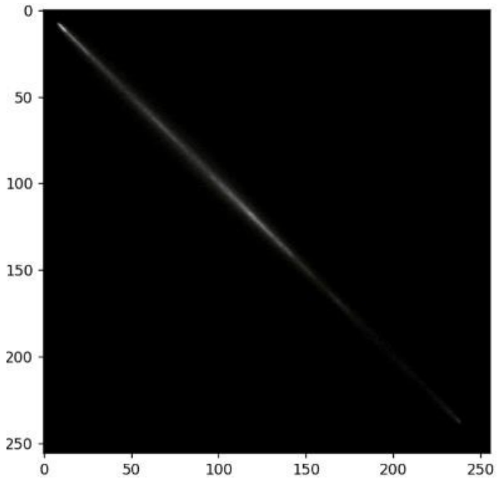


Figure 40: Blurred Version of Figure 39



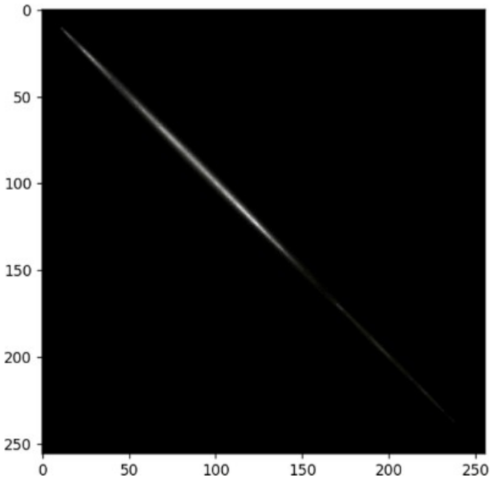


Figure 41: Gray Level Co-occurrence Matrix of Figure 39

Figure 42: Gray Level Co-occurrence Matrix of Figure 40

After generating the Gray Level Co-occurrence Matrix (GLCM), five statistical measures can be calculated from it to help us extract texture features. Let p (i ,j) represent the value of the pixel pair (i , j) in the GLCM.

I. contrast

Reflects the clarity of the image as well as the depth of the texture grooves. The deeper the texture, the greater the contrast, and the clearer the visual effect; conversely, the smaller the contrast, the shallower the texture grooves, and the more blurred the visual effect.

$$Contrast = \sum_{i,j}(i-j)^2 p(i,j)\ ;\ \ i,j = 0,1...255$$

II.correlation

Calculates the degree of similarity in the grayscale values of the image in both the row and column directions. The greater the degree of similarity, the clearer the visual appearance, and hence the higher the correlation. If there is a large difference in grayscale values within the gray level co-occurrence matrix, then the correlation value will be lower.

$$Corr = \sum_{i,j} \frac{(i-u_i)(j-u_j)p(i,j)}{\sigma_i \sigma_j}\ ;\ \ i,j = 0,1...255$$

III.energy

Reflects the uniformity of the grayscale distribution and the fineness of the texture in an image. If the values of the elements in the gray level co-occurrence matrix are distributed uniformly, then the energy is lower, indicating a finer texture; if the values of its elements are unevenly distributed and have significant differences, then the energy is higher, indicating that the original image's texture has a certain degree of regular variation.

$$Energy = \sum_{i,j} p(i,j)^2 \ ; \ \ i,j = 0,1\ldots255$$

IV.Dissimilarity

A measure of the distance between areas of interest (pixels) used to assess the degree of texture irregularity. Higher dissimilarity indicates significant irregular texture areas; lower dissimilarity means the texture areas are more uniform and similar.

$$Dissimilarity = \sum_{i,j} p(i,j) * |i-j| \ ; \ \ i,j = 0,1\ldots255$$

V.homogeneity

Measures the homogeneity of texture, reflecting the degree of local variation in texture. A higher value indicates less variation and more uniformity across different areas of the texture, meaning higher texture smoothness; lower homogeneity indicates greater local variation in texture and lower texture smoothness.

$$Homogeneity = \sum_{i,j} \frac{p(i,j)}{1+|i-j|} ; \ \ i,j = 0,1\ldots255$$

Since there are four angles, one image will have four gray level co-occurrence matrices, one for each angle. Therefore, there will be a total of 20 statistical measures.

D) Correlation Coefficient Chart of All Variables Before and After Dimension Reduction

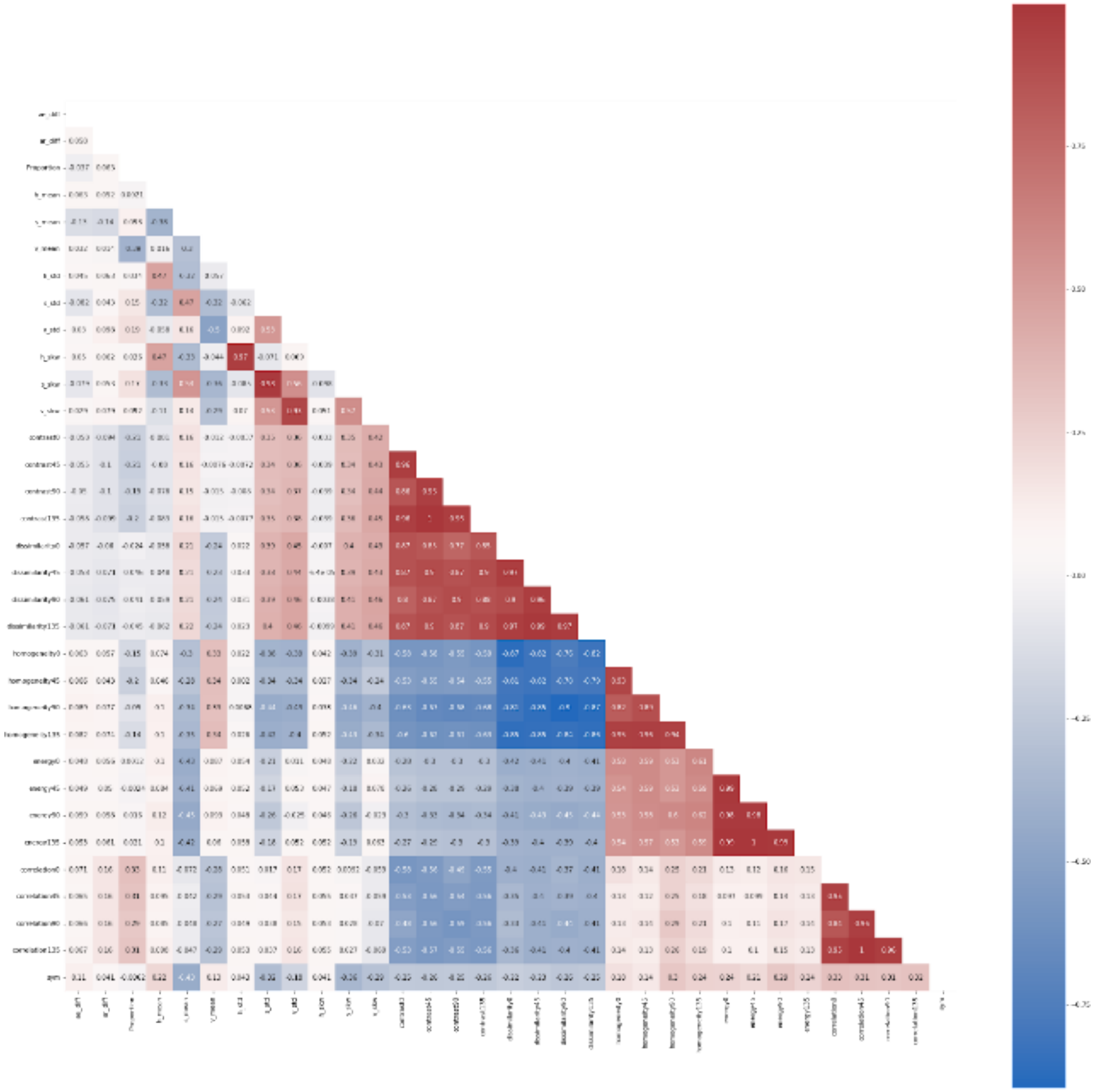


Figure 43: Correlation Coefficient Chart of All Variables Before Dimension Reduction

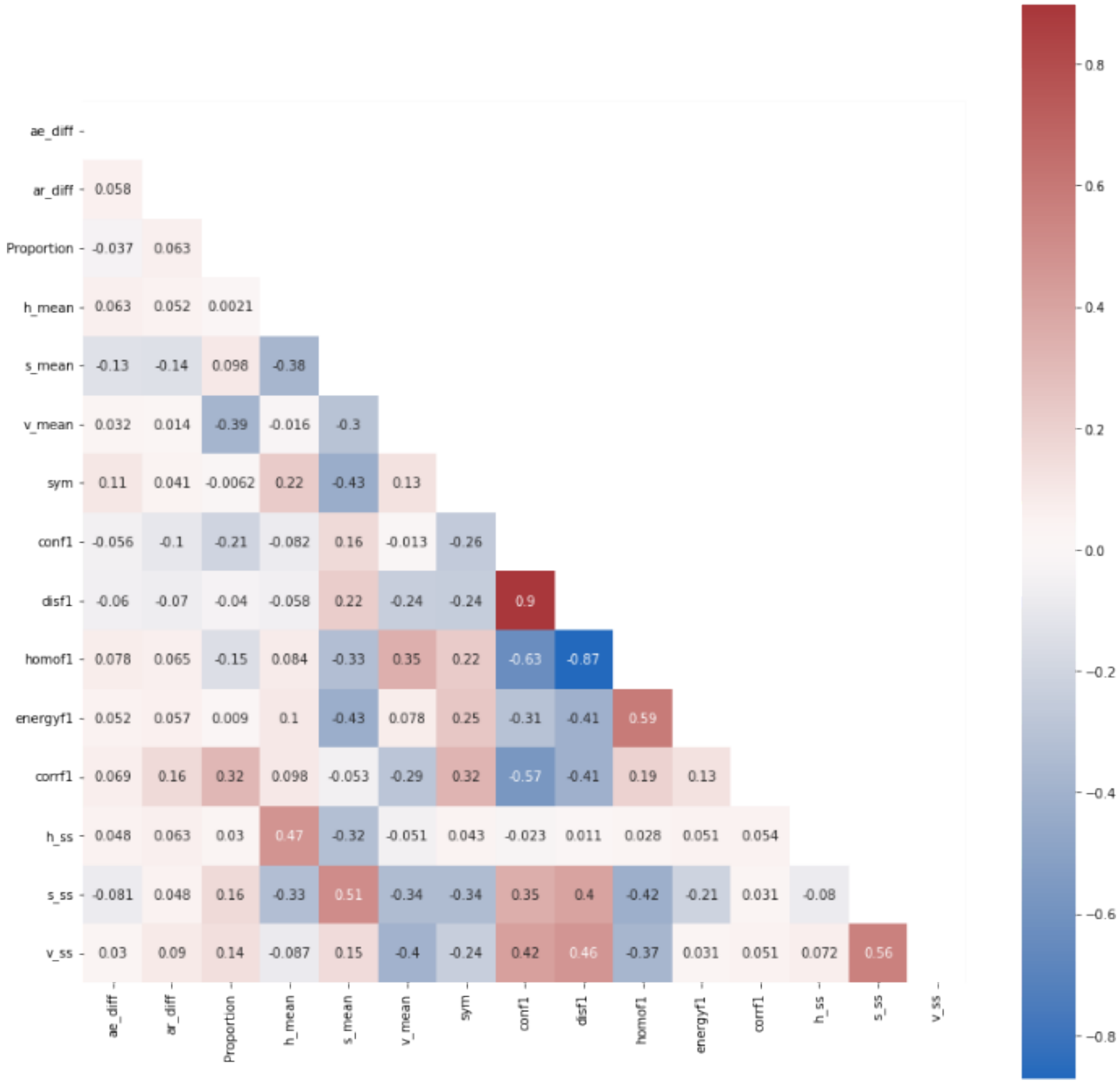


Figure 44: Correlation Coefficient Chart of All Variables After Dimension Reduction

E) Model Introduction

I. Support Vector Machine (SVM)

Support Vector Machine (SVM) is a type of supervised learning that involves mapping data points in a high-dimensional space and constructing a hyperplane for classification, regression, or other tasks. Given a set of training data, each data point is labeled as belonging to one of two categories. SVM assumes that there exists a hyperplane $\vec{w}^{\mathrm{T}}x + \vec{b} = 0$ that can perfectly separate the two sets of data. By calculating $\vec{w}$ and $\vec{b}$,the goal is to maximize the margin between the two categories and reclassify the data into one of the two categories. It is important to note that standardization of each feature in the data is necessary before applying SVM, to ensure that each feature is equally important in the SVM.

II.Random forest

Belongs to Ensemble Learning methods. When training classifiers, multiple decision trees are constructed, and bootstrap sampling is used for training. The final classification result is determined by majority vote. The branching of the trees is based on entropy or Gini impurity to find the variable within the feature subset that has the maximum information gain (IG), thereby stopping the splitting of nodes. Compared to a single decision tree, this method can reduce bias and variance more effectively and is less prone to overfitting.

III.Naive bayes

It is a method for constructing classifiers. It is not a single algorithm for training such classifiers but a series of algorithms based on the same principle: all Naive Bayes classifiers assume that the features of the samples are independent of each other. In many practical applications, the parameter estimation for the Naive Bayes model uses the maximum likelihood estimation method. In other words, the Naive Bayes model can be effective even without using Bayesian probabilities or any Bayesian models. Despite the simplicity of the assumption, Naive Bayes classifiers can achieve quite good performance in many complex real-world situations. The advantage is that it only requires a small amount of training data to estimate the necessary parameters (mean and variance of the variables), and because of the assumption of variable independence, it is not necessary to determine the entire covariance matrix.

IV.K-Nearest Neighbors Algorithm (KNN)

Also known as the KNN algorithm, it belongs to supervised learning. The training data consists of labeled data, and the Euclidean distance between each sample point and the target point is calculated. The k closest points are selected for class determination. There are two methods of determination: voting and weighted voting, with the classification result being the category with the most members. The difference between the two is that the latter considers the distance of the neighbors for weighting before classification. The advantages include high accuracy, no assumptions about data input, and no restrictions on data types. The disadvantages are high time and space complexity, and a high dependency on sample balance.

V.Logistic Regression

Primarily focuses on exploring the relationship between dependent variables and independent variables, differing from ordinary linear regression in that the dependent variable is categorical, especially when divided into two categories. The advantages include ease of calculation and training, and there's no need for the assumption of normal distribution for independent variables, hence fewer restrictions.

VI.Neural Network (NN)

It is a type of nonlinear statistical model that uses neurons to transmit signals, importing information from the input layer into an artificial neural network. Through the nonlinear transformations in the hidden layers and after passing through multiple hidden layers, the output layer ultimately provides the predicted response variable of the artificial neural network.

VII.Convolutional Neural Network (CNN)

Compared to neural networks, Convolutional Neural Networks (CNNs) have the advantage of shared weights, utilizing a convolutional kernel composed of the same few neurons to learn identical features, thus achieving parameter efficiency. CNNs begin with the convolutional kernels sliding over the image to extract information, and then control the dimensions of the image through strides and padding. Subsequently, through pooling layers, the image is compressed while retaining essential information. Afterward, flattening is employed to bridge the CNN layers with the fully connected layers, where feature extraction is conducted in the final fully connected layers.

7. References

1. American Cancer Society - Key Statistics for Melanoma Skin Cancer (https://www.cancer.org/cancer/types/melanoma-skin-cancer/about/key-statistics.html)

2. American Academy of Dermatology Association (https://www.aad.org/media/stats-skin-cancer)

3. Majumder, S. and Ullah, M. A. (2019). Feature extraction from dermoscopy images for melanoma diagnosis., SN Applied Science, vol. 1, 753.

4. 余佳蓉 (2021)。智慧醫療領域對黑色素瘤的生物醫學影像分析。明志科技大學電機工程系碩士論文。

5. Gessert, N., Nielsen, M., Shaikh, M., Werner, R. and Schlaefer, A. (2020). Skin lesion classification using ensembles of multi-resolution EfficientNets with meta data. MethodsX, 7, 100864.

6. Rehman, M., Ali M., Obayya M., Asghar J., Hussain L., K. Nour M., et al. (2022). Machine learning based skin lesion segmentation method with novel borders and hair removal techniques. PLoS ONE, 17 (11), e0275781.

7. ADDI Project. (2013). PH² Database. [Data file]. (https://www.fc.up.pt/addi/ph2%20database.html)

8. Giotis, N. Molders, S. Land, M. Biehl, M.F. Jonkman and N. Petkov: "MED-NODE: A computer-assisted melanoma diagnosis system using non-dermoscopic images", Expert Systems with Applications, 42 (2015), 6578-6585.

9. ISIC Challenge. (2018). ISIC Challenge Datasets. [Data file]. (https://challenge.isic-archive.com/data/#2018)

10. Kaggle. (2020). SIIM-ISIC Melanoma Classification. [Data file]. (https://www.kaggle.com/competitions/siim-isic-melanoma-classification/data)

11. OpenCV. (n.d.). Retrieved May 27, 2023, from https://docs.opencv.org/3.4/db/df6/tutorial_erosion_dilatation.html

12. OpenCV. (n.d.). Retrieved May 27, 2023, from

https://docs.opencv.org/4.x/d7/d4d/tutorial_py_thresholding.html

13. Murzova, A. and Seth, S. (2020). Otsu's Thresholding with OpenCV. Retrieved May 27, 2023, from https://learnopencv.com/otsu-thresholding-with-opencv/

14. Edubirdie. (2022). Skin Cancer: Convolutional Neural Network Based Skin Lesion. Retrieved May 28, 2023, from https://edubirdie.com/examples/skin-cancer-convolutional-neural-network-based-skin-lesion/

Project Competition Report Task Allocation Table

Student ID	Name	Division of Labor
410978004	郭依璇 (YI-HSUAN KUO)	Data and literature collection, process conceptualization and arrangement, program design and development, report writing, and PowerPoint presentation creation.
410978011	譚靖蓉	Data and literature collection, process conceptualization and arrangement, program design and development, report writing, and PowerPoint presentation creation.
410978044	謝瑋芸	Data and literature collection, process conceptualization and arrangement, program design and development, report writing, and PowerPoint presentation creation.
410978056	黃名揚	Data and literature collection, process conceptualization and arrangement, program design and development, report writing, and PowerPoint presentation creation.
410978058	李博業	Data and literature collection, process conceptualization and arrangement, program design and development, report writing, and PowerPoint presentation creation.